Assessing Writing xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Reflecting on the past, embracing the future

Liz Hamp-Lyons

Centre for Research in English Language Learning and Assessment, University of Bedfordshire, Putteridge Bury, Hitchin Road, Luton, Bedfordshire LU2 8LE, UK

In the Call for Papers for this anniversary volume of *Assessing Writing,* the Editors described the goal as "to trace the evolution of ideas, questions, and concerns that are key to our field, to explain their relevance in the present, and to look forward by exploring how these might be addressed in the future" and they asked me to contribute my thoughts.

As the Editor of *Assessing Writing* between 2002 and 2017—a fifteen-year period—I realised from the outset that this was a very ambitious goal, one that no single paper could accomplish. Nevertheless, it seemed to me an opportunity to reflect on my own experiences as Editor, and through some of those experiences, offer a small insight into what this journal has done (and not done) to contribute to the debate about the "ideas, questions and concerns"; but also, to suggest some areas that would benefit from more questioning and thinking in the future. Despite the challenges of the task, I am very grateful to current Editors Martin East and David Slomp for the opportunity to reflect on these 25 years and to view them, in part, through the lens provided by the five articles appearing in this anniversary volume.

Ed White, the father of the modern field of writing assessment in the United States, had the first article in the very first issue of *Assessing Writing*. How fitting it is that he should have the first article in the anniversary volume. Already in 1994, White found it necessary to begin: "As we move to the end of this century, the issues and problems in writing assessment have shifted and become more complicated." (Vol 1: 11–27, 1994) He acknowledges in this new paper that those early issues of *Assessing Writing* focused mainly on teachers' views and needs, with most other interest groups receiving little attention in this field and in this journal. That was true: but such was the context at that time that the apparent privileging of teachers' voices was a much-needed balancing. That was the time (as I well remember) when in the US the 'big tests' held sway, and my American colleagues had been fighting for more humanistic forms of assessing learners' ability to write for many years, because in the US it had become common "to think of the direct assessment of writing as a reaction against and result of multiple-choice, so-called 'objective' testing." (Hamp-Lyons, 2002: 6) But as a Brit, despite having worked for ten years in US composition and applied linguistics programmes, I was less familiar than many or even most of my readers with writing assessment practices in the US. While educating myself in the US writing assessment 'community', I also consciously attempted to locate modern work in writing assessment, which was almost entirely US-based, within a much larger context. I began with a very short history of writing assessment. I showed that the assessment of writing had begun in Imperial China, and over several thousand years (depending on the criteria one adopts), it had spread slowly, first through Europe, then back to South and East Asia as the British Empire spread, before arriving in the US, notably via Harvard University's 1873-4 introduction of a written composition as an entrance examination, replacing the traditional oral examination (Lunsford, 1986).[1]

As White importantly reminds us in his article in this volume, as the field of educational measurement grew and strengthened, it remained, and still remains the case that teachers of writing continue to teach and to provide their students with feedback. However, it must be said that this journal has not, as whole, focused on what students want: it has done rather better, in my view, at focussing on what students need. When I took over the Editorship of this journal in 2002, I saw the future readership of the journal as, first, writing assessment practitioners who also worked as writing teachers and/or in institution-based writing assessment programmes or writing centres. This focus was in line with the tradition that the journal had built up under Editors Huot and Yancey. It was not until Vol 9, No. 2 that *Assessing Writing* published its first paper from a major test development and supplier company (Hawkey and Barker,

E-mail address: Liz.Hamp-Lyons@beds.ac.uk.

[1] A brief history of assessment for education and work can be found at http://www.nea.org/home/66139.htm.

https://doi.org/10.1016/j.asw.2019.100423

of Cambridge Assessment, UK). In Vol. 10, No. 1, the journal published a paper from a research study sponsored by ETS but primarily authored by academics (Cumming et al.) and another paper from a colleague at Cambridge Assessment (Green). Since then, papers by writing assessment researchers in testing organizations as well as in large university assessment programmes, and smaller research units have appeared. But my own glance through the 40+ volumes of the journal suggests to me that assessments in the contexts of learners are still very frequently the concern of our authors.

In addition to seeking to present the story of writing assessment as much longer than had typically been seen in *Assessing Writing,* I also sought to present it as much wider. Writing is assessed everywhere these days; writing in English is assessed almost everywhere too. Therefore, when I became Editor, I wrote in my Introduction:

> *Assessing Writing: An International Journal* is going to publish articles from anywhere (in the end, I hope, from everywhere!) in the world; by nonnative and native writers of English; by scholars, researchers, teachers at all levels of education, test developers and administrators; from those who believe in tests and by those who don't. *Assessing Writing: An International Journal* is going to take the broadest possible view of "writing" and its assessment, constrained only by the submissions we receive." (Volume 8, Number 1: 1–2)

As a reader of the journal from the first issue, and as an original Editorial Board member, I had at first been unaware of my unusual status as the only non-American on the Board: I had just moved from the University of Michigan to the University of Colorado in Denver so my foreignness, my Britishness, did not stand out. But by the time Ablex had sold the journal to Elsevier and Elsevier had asked me to take on the Editorship, I had been working in Hong Kong for five years and was much more conscious of the US orientation of the journal. I had been aware of the limitations of the journal as an international outlet and educational resource. Ablex was an excellent but small US publisher, and very hard to get hold of beyond North America. I wanted all the people I knew who were doing good work in assessing writing beyond North America to be able to read the best, most up to date, research: and I wanted this research to become accessible, and read, in Europe, Asia, New Zealand and Australia, South America and the continent of Africa. I had not forgotten that before arriving in the US in 1986 fresh from a PhD in language testing at the University of Edinburgh, I had studied a very 'foreign' field at a time when not even an august institution like Edinburgh had access to foreign materials in minor fields such as applied linguistics, at a time before online-anything, and before buying books from overseas was technically or financially feasible (and in fact it is still often not feasible, with prices often around US$100 for an academic book). I read as much as I could get hold of in the US literature specifically on the assessment of writing. I survived on inter-Library loans and often-illegible microfilms: often when these texts finally arrived I had only 48 hours to read and make notes from them. I had known nothing of NCTE journals, which even today are not accessible through normal web access without a membership of the NCTE organisation (currently $50 a year plus $25 for each journal). Not surprising, perhaps, that I knew quite a lot, but there were deep and wide holes in my knowledge.

The basis of my knowledge about how people had assessed writing came primarily from British sources (e.g., Edgeworth, 1888; Britton, 1963; Finlayson, 1951; Hartog, Rhodes, & Burt, 1936; Vernon & Millican 1954; Wiseman, 1949). There was indeed plenty of statistical material there; but there was also plenty of worrying about not just the reliability of the essay exam but what we would now call its validity and its educational consequences. After 1986, as I dug into the US literature with the privilege of the University of Michigan Library and an NCTE membership, I was very struck by the strength of the bifurcation between assessment of 'writing' seen as composition, and the measurement of written text seen as a test task type. My own dissertation had prepared me to work with Classical Test Theory, but the US psychometric tradition appeared to be more interested in validating methods than in investigating student performance. There was a technical sophistication that was admirable, but I often felt as if the student had somehow gotten lost in the elegant equations and sophisticated designs.

The university sector in the United States of America has a strong tradition of both developing and critiquing writing assessments of many kinds (Hamp-Lyons 2002; Huot, 2010; Yancey, 1999). Yet this tradition has been primarily qualitative, and heavily dependent on narrative/experiential reporting and interpretation. However, outside the universities and colleges, things were different. Because the introduction of a written composition as an entrance examination, replacing the tradition oral examination, occurred at close to the same time as the surge of research into and use of statistical tools for intellectual inquiry in many fields, and because over time the number of pieces of writing that needed to be assessed grew so large, statistical tools were soon brought to bear on this perceived problem. These tools revealed considerable unreliability between essay markers leading in turn to a focus on standardization through so-called 'objective' tests (Lehmann, in *The Big Test*, and Spolsky, in *Measured Words*, discuss some of the social, and perhaps ethical, consequences of the great love affair with statistics, as well as the growth of 'big testing'). Work on methods of standardization, and on ways of achieving that standardization (e.g., Edgeworth, 1888; Hillegas, 1912; Odell, 1928; Thomas, 1931; Stalnaker, 1936) seemed by the 1960s to be dominating concerns about writing assessment. Early studies into essay marking using human raters, such as that by Diederich (whose "professional love" Haswell (2014) notes, "was ever high school life") and colleagues at Educational Testing Service (Diederich, French, & Carlton, 1961), offered discouraging results on the potential reliability of essay scorers, although from a modern perspective on the scoring of writing their study had many flaws.[2] But by the time I arrived in the US, the voices of writing instructors and writing assessment practitioners were becoming louder, facilitated especially through the NCTE organisation Conference on College Composition and Communication and NCTE's College-focused journal, College Composition and Communication. The gulf between this community and the large-scale test development and delivery organisations seemed at the time (to me) to be growing wider. But in the ten years I worked in the US between 1986 and 1996, first at Michigan and then at

[2] A new book by Richard Haswell and Norbert Elliot, on *The Early History of Holistic Scoring*, is due out in November 2019.

Colorado, I watched change slowly taking place in the educational measurement literature. The work of Cronbach (1988), Madaus (1988) and Messick (1988, 1989) with its emphasis on ethics, consequences and validity resonated with my British roots and training (Hamp-Lyons, 1997b). Slowly an awareness of the relevance of some work in educational measurement trickled through into the writing assessment literature: this was undoubtedly helped when ETS introduced the Test of Written English to the TOEFL, and ETS researchers put more efforts into research into the assessment of writing. In my time at Michigan I had the privilege of working with my colleague Bill Condon to introduce portfolio assessment to the composition programme at Michigan (Condon & Hamp-Lyons, 1991).

By 2001, when I was invited to take on Editorship of *Assessing Writing*, this opportunity spoke to several issues that were occupying my mind, and I accepted the challenge. In articulating my goals for this journal I talked about "[raising] consciousness of the political and ethical dimensions of the work we do, in our classrooms, our school districts/systems, our colleges and universities, and our countries." I didn't use the word 'fairness', but I might have done, if I had not seen 'ethics' as inclusive of and superordinate to fairness.

Messick (1988, 1989) had already argued that "the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force behind this integration is the trustworthiness of the empirically grounded core interpretation, that is, construct validity" (1989, p5). Messick's thinking has stimulated substantial further work; Kane in particular has emphasized the need to make arguments for claims about test validity (Kane, 2013). Closely related to some of the concerns of writing assessment, the work of Stephen Toulmin (1958, 2001) has been applied and extended in the creation of what is now generally called an interpretation/use argument (IUA) Bachman (2005); Olivieri, Lawless, & Young (2015). In relating his work to the field of language testing, Messick (1996) addressed *washback*, a term which, as he says, refers to "the extent to which the test influences language teachers and learners to do things they would not otherwise do" that promote or inhibit language learning" (p. 241). This and other work in the field of language testing showing the important link between construct validity and washback, seen in Messickian terms, has helped us to better understand the ethical aspects of what we do, and has provided a rich body of empirical research on washback (Cheng, 2014; Hamp-Lyons, 1997a).

In my first issue as Editor, I contributed a short article of my own which gave me the opportunity to give my personal perspective on 'the scope of writing assessment'. I argued that the future for writing assessment must be at once technological and humanistic, harnessing the power of computing for large-scale assessments together with its increasing subtleties for affording more options for writers, raters and teachers. To some extent this has happened, but more in relation to testing agencies than to individual learners, teachers, raters and decision makers within educational settings. I also argued that "the political nature of all assessment must be realised" and that we need to understand that "what we do when we design or administer a writing test, when we score it, when we take and utilise test scores, is to participate in a form of social engineering that is at once beneficial and dangerous" (p13).

Where, then, do I think that the field has gone and is going, after 40 years or so and after 25 years of *Assessing Writing*?

Towards the end of his paper in this volume, White asks: "What kind of writing assessment, then, do students want?" He answers his own question based on recent experience reading students' writing in a MOOC (Comer & White, 2016).

- Assessment designed to provide maximum, actionable, and speedy feedback to the writer
- Assessment that breaks down the complexity of writing into focused units that can be learned in sequence and improved by practice
- Assessment that produces data principally for the use of teachers and learners
- Assessment that focuses on critical thinking and creativity and that places surface features of dialect and usage in a large social context

These priorities will not necessarily be the same for learners in other contexts, but the importance placed on feedback by the students in Cromer & White's study adds to the now overwhelming evidence that writers at every age and in every context appreciate sensitive feedback from trusted peers and experienced teachers who empathise with the task of creating text which is now coming from around the world. Likewise, evidence that students, their teachers and their parents desire data from writing assessments that is **useful to them** seems indisputable. It is less clear what large-scale testing companies are doing to provide this, beyond the creation of fairly simple technological applications such as Criterion (ETS) and MYAccess! (Vantage Learning). Stevenson and Phatiki (2014) found only "modest evidence" that AWE feedback resulted in improvements to the texts that received AWE input, and no evidence of general improvement in writing; Stevenson (2016) reported a critical interpretive synthesis of AWE software (Summary Street, MY Access! and Criterion) and concluded that "Surprisingly few criticisms have been leveled against the ability of AWE to develop students' revising skills, which surely should be a central objective of programs that provide students with multiples drafting opportunities and detailed automated feedback." (p.12) An excellent paper by Elliot et al. (2013) provides a detailed critical bibliography of articles and book chapters on AWE software. However, referring to such software, Zhang & Hyland (2018) make the point that "the mere provision of feedback does not automatically lead to writing improvement. Rather, it is the effective student engagement with this response to their work that is likely to unlock the benefits of feedback." (90). The significant numbers of papers relevant to feedback that have been published in *Assessing Writing* over the years, in contrast, has been a great resource for readers, and has made clear that, in the current period, we regard feedback as a key element in a humanistic model of 'assessment'.

When later in his paper. White turns to "what testing organizations (and governing bodies) want from a writing assessment" (in press), I find I can agree with him only in part. He is certainly correct that "testing firms want" assessments that produce scores and other data quickly and inexpensively, but in my view testing firms don't want their tests to "reduce the complexity of writing and the teaching of writing yet allows users to make [construct-valid inferences]", or to "depend heavily on statistical explanations of

sufficient complexity to allow justified inferences". These are outcomes partly of the technology taking an assessment to a large scale, and of demanding fast turn-around of results: they, seemingly inevitably, arise from the culture of testing and its marriage to bureaucracies. But my own experience and my reading of the research literature published from two major language testing organizations with enormous candidatures of English as a second language users, Educational Testing Service (US) and Cambridge Assessment (UK and Australia) encourages me that these organizations have taken on modern views not only of the psychometric aspects of tests and other instruments of assessment, but are giving increasing attention to fairness defined more inclusively than DIF and bias studies, and moving towards thinking about issues of social justice and OTL (opportunity to learn). This can be seen in their own online Research Reports series, as well as publications in a wide range of scholarly journals. However, research into national exams in the UK, such as GCSE (General Certificate of Secondary Education), and in Australia into the major schools' examination NAPLAN (National Assessment Program – Literacy and Numeracy) is more difficult to find. My own involvement in a review of the NAPLAN for its official body ACARA (Australian Curriculum, Assessment and Reporting Authority) was not a salutary experience (Hamp-Lyons & Wolfe, 2018) and Perelman (2018) tells a parallel story in a study sponsored by the Australian Federation of Teachers (Perelman, 2018). In this area, assessment bodies in the US do a much better job than those in many other countries.

An area which I was very involved in on a personal professional level when I took on Editorship, but failed to emphasize, except, perhaps, implicitly, in my introductory paper, was the need for 'assessment instruction', meaning training and instruction in writing assessment methods, practices and quality issues as well as the training of raters and teachers. In the US, most composition teachers in colleges and universities are privileged to have received instruction in teaching writing in high quality graduate programmes: this is largely explained by the fact that 'First Year Writing' in the US requires a large number of relatively inexpensive teachers, while graduate students require a means of paying their way through college. But in the rest of the world this kind of thing is almost unknown. Even in the US, until quite recently it has seemed that assessment of writing received little attention in taught coursework (there being some outstanding exceptions). Generally, and even in programmes of training for writing teachers, writing assessment research happens, if at all, only at doctoral level. A result (as is shown by Zheng and Yu, this volume) is the lack of articles about training and instruction in writing assessment methods, practices and quality issues. From my perspective, still today, when beginning writing teachers encounter assessment practices that confuse or worry them, they may well turn to White's seminal book (first published in 1985, second edition 1994) for advice and clear examples, in addition to Crusan's much more recent book (2010).

White, in his paper, goes to comment that "Fairness, though related to validity, is not just a colloquial relief from statistical jargon, but a distinct and measurable aspect of assessment" (in press). The question whether 'fairness' is measurable is an interesting one, and is addressed in this volume by Poe and Elliot. Kane (2010), responding to a very thoughtful paper by Xi (2010) says:

> Validity theory has tended to focus on the accuracy and appropriate-ness of score-based interpretations and decisions about all of the individuals in the population of interest. Analyses of fairness have tended to focus on group differences and on differences in the accuracy and appropriateness of interpretations and decisions across groups, which are defined in terms of race/ethnicity, gender, age, and so on. (p. 181)

The question of how far measuring accuracy and appropriacy relative to different groups goes as a solution to issues of fairness is important and significant, as indeed White points out. Kane (2010) discusses a care originally raised by Shepard (1993), who, referring to a kindergarten readiness test, argued that even when a standardized 'readiness' test is shown to be an accurate measure of certain basic skills, and is also found to be an excellent predictor of performance in kindergarten, decisions about readiness still deserve more evidence. If other evidence found low 'readiness' to be caused in some cases by a developmental lag, waiting a year might seem a sound strategy. However, if richer data could uncover, in some cases, for example, a home environment that does not support the child's development of the necessary skills (i.e. a lack of opportunity to learn), keeping the child out of school for another year may well be. Such variation from child to child may well be due to, often invisible, or at least unobserved, 'disparate impact'.

'Disparate impact' is raised in the paper by Poe and Elliot in this volume, the title of which is "Evidence of fairness." They began with a key word search on the word "fairness" and then expanded to include the terms bias, ethics, and justice. They read the papers to check how the term had been used by the paper's author(s), and when "some articles used a term, such as "fairness," generically… These articles were not included for further study because they were not grounded in a theoretical orientation to the use of the term." Poe and Elliot acknowledge that each of these Key Words may mean different things to different people and, indeed, to different groups (what in genre analysis get called 'discourse communities'). Indeed, I soon understood that my understanding of 'fairness' was not the same as theirs. Their first question asks: *How have writing assessment researchers constructed fairness?* and in explaining a second stage to their study they discuss 'fairness' from three perspectives, a 'generational'approach; a focus on "U.S. educational measurement standards as they have changed from 1952 to 2011"; and through "changes [to] the way[s] researchers have used evidence related to fairness, such as in bias research". I am aware, and rather unapologetic that, my understanding of fairness as an assessment principle is grounded in an orientation very unlike the one they share. This is perfectly normal and completely acceptable: but I think that a different orientation would have yielded rather more encouraging results than their analyses led to.

From my point of view, using the US standards in this international journal necessarily limits the Authors' perspective on definitions of 'fairness' to a single national view; and the close linking of 'fairness' with 'bias', a term used within a specific statistical approach that, while somewhat related to fairness, and more distantly related to justice and ethics, is similarly tied to the US standards. The reasoning behind this becomes clearer as, after a short history of the journal, Poe and Elliot introduce a quote from Dorans (2011) in which he summarizes his view of four generations of fairness research, and describes the fourth generation as "recogniz[ing] that testing occurs within a larger complex system and that measurement needs to occur within this larger context." Poe and Elliot then say "Here we see the evolution of fairness as it has evolved… Overall movement extends from a technical view of

fairness (as bias) to an embedded view of fairness (as situated within contexts)."[3] This identifying of an "overall movement" in this positive direction is optimistic, and to some extent I can see it in the field and in *Assessing Writing*. But I think if Poe and Elliot had included those papers that "used the term "fairness," generically… [and] were not grounded in a theoretical orientation to the use of the term" they might have found some which had valuable insights for us in building a better argument for 'fairness' in writing assessment development, assessment uses, and research, both in the US and elsewhere. Poe and Elliot add that "In later work, Dorans (2017) has provided a review with research associated with each generational phase, save the earliest". However, my own reading of the Dorans paper found only the mildest of caveats about the value and meaning of DIF as a fairness indicator. In contrast, writers such as Lay & Papadopoulos (2007), in the UK citing Guba and Lincoln (1989), and Elder (1997) and McNamara, Knoch and Fan (2019) in Australia, have probed other possibilities. Lay & Papadopoulos (2007) describe fourth generation evaluation (FGE) as constructivist evaluation, in which "the fairness criterion [of FGE] is designed to ensure the equal participation of disenfranchised and disempowered groups, while protecting them from exploitation. Fairness is also measured by 'the extent to which different constructions and their underlying value structures are solicited and honoured within the evaluation process" (Lay & Papadopoulos, p.486). Lay and Papadopoulos also comment that "being fair may be one of the most difficult of the FGE criteria to meet". I was very pleased to see that Poe & Elliot singled out Petersen (2009): this paper reminded me of a minor finding of my own PhD: 'resistant' writers were rare, but their writing tended to be scored well above or well below the normal curve; I hypothesized that this was explained by writer-reader interactions, and wondered whether in such cases our judgements might be unfair, that is, might fail to correctly judge the writing capabilities of these individuals. I would greatly welcome more attention to such an 'embedded' view of fairness, the view promoted by Guba and Lincoln, and which has been a key part of my own work for many years (see Hamp-Lyons, 2001a, 2001b; 2007).

My sense is that many *ASW* papers address their research issues from an embedded awareness of what I would think of as 'fairness'. Let me nominate one paper, almost at random from my last few issues as Editor: Sadeghi and Talmahi (2017) conducted an empirical investigation in which they introduced an integrated assessment built on assessment *as*, *for* and *of* learning (a three term-modifications becoming increasing used in the assessment for learning literature). They found that the integrated assessment types model showed some benefits, and they emphasized several ways in which they supported learners in achieving success using the three forms of assessment, particularly stressing the value of "offering conceptualizations of self-assessment, self-evaluation, and self-reflection" and "making use of standard and detailed assessment criteria" (p.60). For me, this attention to what teacher-researchers can do to help writers achieve their very best is a key part of an approach to writing assessment research, or any other research, that is aimed at ensuring that every learner will be the best they can be. It is indeed true, as Poe and Elliot, citing Dorans (2017), say, that "Not all fairness considerations can be reduced to quantitative evaluations" (in press). In another paper Poe and Cogan (2016) discuss complex issues of US Civil Rights legislation (that is well beyond my knowledge and world view), but in which I was heartened by their quotation from President John F. Kennedy:

> "not every child has an equal talent or an equal ability or equal motivation, but they should have the equal right to develop their talent and their ability and their motivation, to make something of themselves." June 11, 1963)

However, as Poe & Cogan went on to describe 'disparate impact', it seems to be a uniquely American concept, a mostly legal/technical perspective on fairness, again dependent on forms of data disaggregation. From what I can understand of the vast weight of legal demands and prohibitions faced by assessment researchers once they step beyond their academic doors, this emphasis is probably not surprising. But we need to go further: the International Language Testing Association (ILTA) echoes President Kennedy with one of its mottos: "not everything that can be counted counts, and not everything that counts can be counted" (attribution uncertain).

For example, Slomp (2016) usefully summarizes his study of the Alberta English 30-1writing, in which he found that applying Kane's version of the IUA led indisputably to the conclusion that "Even under the most cursory of reflections, it is painfully obvious that no single-sitting, timed, impromptu writing task, no matter how well constructed, could ever measure such a broad range of complex learning outcomes and their associated constructs." He found that "teachers narrowed their marking criteria, the number of genres they assigned, and the range of curriculum outcomes they assessed to the narrow set of criteria, tasks, and outcomes measured by the diploma exam" and that "students learned to value the narrow set of skills measured by the diploma exam in their own development as writers". Findings like these mirror the very important issues that Smith (1991) notably wrote about, and in language testing we too have been writing about them under the name of 'washback' for 25 years (Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996). In rigid test contexts, some students shine and some crash and burn. Here is a clear instance where the disaggregation of data that Poe and Elliot argue for could play an important part — but only a part — in research into fairness in writing assessment: the deliberate collection of richer data on all learners, through teachers, parents, school or college advisors, would almost certainly lead to more fairness in decision-making. I have absolutely no argument against the demand for awareness of 'disparate impact', but as defined here it leaves aside many kinds of fairness that the best teachers are aware of and consciously enact in their teaching and classroom assessment practices, and which I believe we find underlying many of the papers published in *ASW*.

The paper by Latif discusses a concept that has received little attention in writing research, and rarely if ever in writing assessment research, although from some perspectives it is discussed often in writing classrooms and among test developers: motivation. I was fascinated to find that he cites almost no writing assessment literature, and only one reference to this journal. This disappoints me,

[3] All quotes from the Poe and Elliot paper in this Volume were drawn from a proof copy I was sent, and may not exactly match the version which will finally be published.

first because his paper convinces me that we, writing assessment scholars and teachers of writing, should pay much more attention than we do to our students' motivation; but second, because I wonder whether, as with Poe and Elliot, and indeed with Hammond, the tight text linguistic approach to data categorization and analysis in the case of such broad concepts may result too narrow—or too broad—an interpretation. The publication of the paper by Wright, Hodges and McTigue in Volume 39 of this journal is, then, fortuitous. Wright et al use a model that is more familiar to me (see their Figure 1, p.74), building from their empirical study to indicate that writing motivation can be seen as having two components, beliefs about self, and beliefs about writing. Their third 'branch' in the model is 'attitudes to writing': I wonder whether in fact attitudes derive from these two kinds of beliefs, implying a somewhat different model. Latif concludes that "The above review shows that writing motivation is an area characterized by a wide variance in conceptualizing and measuring a number of its key constructs." (in press) Perhaps he has created his own problem, to some extent, by accepting such a wide range of terminology as falling within the same sphere. He is certainly right that "Resolving the above-mentioned conceptualization and measurement problems calls for much work at the theoretical and research levels." Having widened our horizons, I hope that he will now take on some of that work.

Zheng and Yu take us back to a more predictable approach to looking back over 25 years. They report on a mainly quantitative analysis of articles in Assessing Writing since 2000: The authors categorised the empirical articles first by the context and participants of the study, then by research focus and theoretical orientation, and then by research methods and data sources. I found this organisation useful, although a far longer treatment would have been needed to unpack all elements of their judgements as they apply for use in establishing a useful separation of measurement theories and writing theories in relation to writing assessment. I was surprised by Table 7, which shows that the emphasis swung gradually more towards quantitative methods and away from qualitative methods in the later period. I suspect this may be because there were more papers from researchers at big testing companies, or contracted to conduct research for big testing companies, in the more recent years, but have not tested my hypothesis; I do feel that mixed methods studies have increased in frequency is not only right, but is in line with the trend in social science research generally. I was intrigued by the datum that there were 7 'psychophysics'-related paper in the second period: I wish I knew which they were! In general, with this paper I felt that I was looking at a useful set of data made ready for close inquiry, but the close inquiry was not in the paper. Doubtless that would have required a much larger study and many more pages of reporting: but it would have been, I think, immensely interesting. My suggestion to the authors, should they feel inclined, would be to address one component at a time within the limitations of a conventional journal article. Some aspects of their data would surely be of interest to journals where researchers' interests intersect with those of *Assessing Writing* readers.

In his paper, Hammond argues that "scholars background racism by eliding or minimizing talk of racial injustice in their writings" and makes it clear that we should be looking at 'race talk' as structural injustice or epistemological racism, a term which refers to "epistemologies, knowledges, and practices that privilege the European modernist White civilization" over other traditions, or at the exclusion of them (Hammond cites Kubota and Lin, p. 479, 2006). In this, he is completely correct, as my other self, dwelling in the world of research into 'English for Academic Purposes', knows very well (2001b, Hamp-Lyons, 2001a; 1997a). The issue I have with his argument (and his very frequent use of the word 'white') is that when he talks about 'race' he clearly has 'blackness' in mind. Yes, there is racism everywhere, including in writing assessment; yes, the tools for judging writing quality are based on a conventional view of "good English" historically derived from the language which is recognizable as "English" as used in Britain since before the time of Shakespeare. Historically this practice derives from British public schools (in Britain a 'public' school is, ironically, an elite private school) and early university entrance practices. The emphasis on knowledge of conventions such as surface features of writing and dialect features of edited Standard (American) English grew beside this tradition. It was integrated into early formal examinations (Weir, 2013) and travelled to the US where much the same structure was introduced to Harvard between the 1860s and 1870s (Brereton, 2012); and my reading of the literature of early examinations of 'English' in the US suggest that this "English" transferred to the former colony with its colonial baggage intact (Newton Scott, 1909; and see Gomes, 2018).

When so-called 'objective' tests began, language proficiency/competency—call it what-you-will, tests included a large proportion of grammar items. Inevitably, these were built on the same model of writing that appeared on written tests and in the textbooks pupils were reading at school and university. For me, there is an inevitability about this (what I think of as) 'cultural imperialism'. But there are other ways of thinking about race. Living, as I have, in several countries and among many cultures during my life, I am much more used to being only peripherally aware of colour and much more aware of language(s), language ability, and how people choose to use language. If I were to think of race as colour, it would be more yellow or brown, as are the peoples of Asia. Though Britain has much to be ashamed of historically in relation to colonialism and indeed slavery, within Britain itself there were few persons 'of colour' until the 1960s. Sadly, with the influx of refugees in the past ten years there is more racism than ever before, but it is often not about colour. Polish, Hungarian and Romanian people are 'white' but are discriminated against in Britain. But they are not discriminated against for their languages, but because of that most difficult prejudice—against 'otherness.'

When I look back at what *Assessing Writing* has been able to achieve, I am proud to see a significant number of articles that have investigated the writing of users of many other languages as they respond to test items, write academic papers, and do other tasks through the medium of English. Present, though perhaps invisibly, are learners of any and all colours as well as many ethnicities, abilities, religions, genders, political persuasions, etc. Authors of these papers have been native speakers of English and second-language users of English, but they have universally been truly concerned for the writers they have worked with and studied.

Conclusions

In the last 25 years, in my view writing assessment has become more technological; we have seen some signs of the technology being used more humanistically, giving more options for writers, raters and teachers, but the promise we saw in the 1980s in

innovations such as computer-mediated peer response tools such as Daedalus has been only partially fulfilled (see, e.g. Peckham, 1996). At that time I also argued that we must understand that all assessment is inherently political with implications not only for large bureaucracies but also for individual learners, teachers, parents and decision makers at every level within education. and that we need to understand that "what we do when we design or administer a writing test, when we score it, when we take and utilise test scores, is to participate in a form of social engineering that is at once beneficial and dangerous" (p13). Nothing I have seen or read since has changed my view of the overtly or covertly political nature of all assessments, including writing assessments. I do see more awareness of the fact in this journal and others, and I hope that awareness is bringing us all more wisdom.

I am pleased to have had, in this reflection, the opportunity to write about several of the issues that continue to occupy my mind as I continue, at my own pace, to consult with writing test developers. I mentioned teacher education in writing assessment earlier: but educating government agencies and test developers around the world is a great need and progress is slow. I've expressed my own views on the importance of fairness as a principle, and of where I think we may find important insights into long-standing issues around lack of, or disparate access to, fairness. I have learned more, and happily, about the increased value attributed to the role of feedback as a process and practice, both in the daily life of the writing classroom and in a more holistic view of writing assessment. But I regret that in the analyses in these papers, the absence of reference to portfolio forms of assessment suggests that they have not taken root as vigorously as many of us hoped they would in the optimistic days of the late 1980s and 1990s, in education, or in this journal. There are many reasons for this, too many to go into here. One of them is the fact that teaching and assessing with portfolios demands teachers with high levels of commitment and knowledge of the principles and methods of teaching and assessing with portfolios. Another is simply because submissions in the area of portfolio assessment have been rare. It should be stressed that a journal can do little to elicit the kinds of papers its Editors and Editorial Board may want. For me, this piece has been an opportunity to reflect on the strengths and weaknesses of this journal, and of the field, in particular the continuing need to do more to ensure fairness, however defined, in writing assessment research and practice.

Several of the papers in this volume seem to have made an assumption that some form of agency at/in/of the journal has made decisions about what kinds of articles to accept. In my Editorship this was never the case and I am sure it is not the case with David Slomp and Martin East. This volume will, I hope, serve as a call to many of our readers to write up their own work in assessing writing and submit it to the journal: we have now seen how badly we need more—more research, more reflective preparation for assessment researchers and practitioners, more reports of writing assessments in all contexts and at all educational levels, more engaging with 'absent' issues such as those raised by Latif and Hammond, more consideration of appropriate methods, as well as some outing of poor practice in ways that can stimulate efforts to make improvements, more overt research into fairness (as clearly defined by the study authors) in real writing assessments, at school as well as college level, and within large-scale testing as well as institutions developing and using their own writing assessments.

Acknowledgement

The author acknowledges the support of the Leverhulme Trust (UK) for an Emeritus Professor grant, No. EM-2018-077-00003.

References

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Brereton, J. (2012). A closer look at the harvard entrance examinations in the 1870s. In N. Elliot, & L. Perelman (Eds.). *Writing assessment in the 21st century: Essays in honor of Edward M. White.* Hampton Press.

Britton, J. N. (1963). Experimental marking of English compositions written by fifteen-year-olds. *Educational Review, 1*(1), 17–23.

Comer, D. K., & White, E. M. (2016). Adventuring into MOOC writing assessment: Challenges, results, and possibilities. *College Composition and Communication, 67*(3), 318–359.

Cheng, L-y. (2014). Consequences, impact and washback. In A. J. Kunnan (Ed.). *The companion to language assessment* (pp. 1130–1146). New York: Wiley & Sons.

Condon, W., & Hamp-Lyons, L. (1991). Introducing a portfolio-based writing assessment: Progress through problems. In P. Belanoff, & M. Dickson (Eds.). *Portfolios: Process and product* (pp. 231–247). Portsmouth, NH: Boynton/Cook.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.). *Test validity* (pp. 3–17). Hillsdale NJ: Erlbaum.

Crusan, D. (2010). *Assessment in the second language writing classroom.* Ann Arbor: University of Michigan Press.

Diederich, P. E., French, J. W., & Carlton, S. (1961). *Factors in judgments of writing ability, research bulletin RB-61-15.* Princeton, NJ: ETS.

Dorans, N. J. (2017). Contributions to the quantitative assessment of item test, and score fairness. In R. E. Bennett, & M. von Davier (Eds.). *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 201–230). .

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599–635.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*(3), 261–277.

Elliot, N., Gere, A. R., Gibson, G., Toth, C., Whithaus, C., & Presswood, A. (2013). *Uses and limitations of automated writing evaluation software. WPA-CompPile research bibliographies, No. 23. WPA-CompPile research bibliographies.* Downloaded 1st Aug 2019 from: http://comppile.org/wpa/bibliographies/Bib23/AutoWritingEvaluation.pdf (Accessed 1 August 2019).

Finlayson, D. S. (1951). The reliability of the marking of essays. *The British Journal of Educational Psychology, 21*(2), 126–134.

Gomes, M. (2018). Writing assessment and responsibility for colonialism. In A. Inoue, & M. Poe (Eds.). *Ch. 6. Race and writing assessment.* Frankfurt: Peter Lang.

Hamp-Lyons, L. (1997a). Exploring bias in essay tests through student interviews. In C. Severino, J. Guerra, & J. Butler (Eds.). *Writing in multicultural settings* (pp. 51–66). NY: Modern Language Association.

Hamp-Lyons, L. (1997b). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295–303.

Hamp-Lyons, L., et al. (2001a). Ethics, fairness(es) and developments in language testing. In C. Elder, A. Brown, L. Grove, K. Hill, N. Iwashita, & T. Lumley (Vol. Eds.), *Experimenting with uncertainty: Essays in honour of alan davies; studies in language testing: 11*, (pp. 222–227). Cambridge: UCLES/Cambridge University Press.

Hamp-Lyons, L. (2001b). Fairnesses in language testing. In A. J. Kunnan (Ed.). *Fairness and validation in language assessment* (pp. 99–104). Cambridge: UCLES/Cambridge University Press SiLT Vol. 9.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*(1), 5–16.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. In J. Cummins, & C. Davison (Vol. Eds.), *The International handbook of English language teaching: Vol. 1*, (pp. 487–504). Norwell, MA: Springer.

Hamp-Lyons, L., & Wolfe, E. (2018). *NAPLAN writing assessment review. Internal report for ACARA* (available from lizhamp-lyons@outlook.com).
Hartog, P. J., Rhodes, E. C., & Burt, C. L. (1936). *The marks of examiners: Being a comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on a viva voce examination.* London: Macmillan.
Haswell, R. (2014). *Paul B. Diederich? Which Paul B. Diederich? Journal of writing assessment 7.1.* http://www.journalofwritingassessment.org/article.php?article=58.
Hillegas, M. B. (1912). A scale for the measurement of quality in English composition by young people. *Teachers College Record, 13*(3), 331–384.
Huot, O.'N. M. (2010). A usable past for writing assessment. *College English, 72*(5), 495–517.
Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
Lehmann, N. (2000). *The big test: The secret history of the american meritocracy.* New York: Farrar, Straus and Giroux.
Lunsford, A. (1986). The past—and future—of writing assessment. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.). *Writing assessment: Issues and strategies* (pp. 1–12). White Plains, NY: Longman.
McNamara, T. F., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment.* Oxford University Press.
Madaus, G. F. (1988). The influence of testing on the curriculum: From compliant servant to dictatorial master. In L. Travers (Ed.). *Critical issues in curriculum: 87th NSSE yearbook, part 1* (pp. 83–121). Chicago: National Society for the Study of Education/University of Chicago Press.
Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Weiner, & H. Braun (Eds.). *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.
Messick, M. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.
Messick, S. (1996). *Validity and washback in language testing. ETS Research Report 96-17*Princeton, NJ: Educational Testing Service.
Odell, C. W. (1928). *Traditional examinations and new-type tests. Englewood Cliffs.* NJ: Prentice-Hall.
Olivieri, M. E., Lawless, R., & Young, J. (2015). *A validity framework for the use and development of exported assessments.* Princeton, NJ: Educational Testing Service.
Peckham, I. (1996). If it ain't broke, why fix it? Disruptive and constructive computer- mediated response group practices. *Computers and Composition, 13*, 327–339.
Perelman, L. (2018). *Towards a new NAPLAN: Testing to the teaching.* NSW Teachers' Federation.
Poe, M., & Cogan, J. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to evaluate social impact. *Journal of Writing Assessment, 9*, 1.
Sadeghi, K., & Talmahi, R. (2017). Integrating assessment as, for, and of learning in a large-scale exam preparation course. *Assessing Writing, 34*, 30–51.
Scott, F. N. (1909). What the West wants in preparatory education. *The School Review, 17*(1), 10–20.
Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Vol. Ed.), *Review of research in education: 19*, (pp. 405–450). Washington, DC: American Educational Research Association.
Spolsky, B. (1995). *Measured words: The development of objective language testing.* Oxford: Oxford University Press.
Stalnaker, J. M. (1936). The measurement of the ability to write. In W. S. Gray (Ed.). *Tests and measurement in higher education* (pp. 203–215). Chicago: University of Chicago.
Thomas, C. S. (1931). *Examining the examinations in English: Report to the college examinations board*Cambridge, Mass: Harvard University Press.
Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition, 42*, 1–16.
Stevenson, M., & Phatiki, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.
Toulmin, S. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.
Toulmin, S. (2001). *Return to reason.* Cambridge, MA: Harvard University Press.
Vernon, P. E., & Millican, G. D. (1954). A further study of the reliability of English essays. *British Journal of Statistical Psychology, 7*(2), 65–74.
Weir, C. (2013). *Measured constructs: A history of Cambridge English language examinations 1913–2012. Cambridge english research notes 51. Cambridge English Language Assessment.* Downloaded 2nd Aug 2019 from:https://www.cambridgeenglish.org/Images/130828-research-notes-51-document.pdf.
Wiseman, S. (1949). The marking of English composition in grammar school selection. *The British Journal of Educational Psychology, 19*(3), 200–209.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.
Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483–503.